IFLA Newspaper Section Pre-Conference
"The Electronic Re-evolution - News Media in the Digital Age"
Mikkeli, Finland, August 7-9, 2012

Preserving News in the Digital Environment: A New Strategy for Research Libraries

Bernard F. Reilly

For centuries, libraries have been the trustworthy custodians of the “first rough draft of history,” as embodied in the published output of the world’s journalists. National libraries and the major academic research libraries created efficient protocols for collecting and preserving the back files of significant newspapers. Their solutions involved workflows tailored to the idiosyncrasies and characteristics of publications issued in printed form on a periodic basis.

The decline of the newspaper industry, however, combined with the ascent of digital media for news reporting and distribution, has up-ended these long-established preservation strategies. Today news production is no longer the periodic, linear process with a single, tangible product that it was in the era of the printed newspaper. Rather news production has become a continuous loop of news gathering, processing, versioning, output, response, and update. Because only a small portion of the daily output of the world’s journalists now appears in printed form, merely acquiring and preserving print editions of newspapers will no longer ensure access to a comprehensive journalistic record for future scholars. Moreover, recent strategies devised to respond to these changes, such as Web harvesting and e-deposit, do an inadequate job of capturing this digital record. I will here discuss where those methods fall short, suggest some alternative strategies, and briefly outline how CRL intends to pursue the latter.

To better understand the landscape of digital news production and distribution we at CRL decided it would be necessary to **map the lifecycle of news content** published in newspapers and online. In a report undertaken for the Library of Congress Office of Strategic Initiatives we laid out the implications of this lifecycle for the strategic preservation of electronic news.

The CRL report “Preserving News in the Digital Environment: Mapping the Newspaper Industry in Transition” grew out of a **workshop convened by the Library of Congress** National Digital Information Infrastructure and Preservation Program (NDIIPP) in September 2009.[1] The workshop explored possible

[1] “Preserving News in the Digital Environment: Mapping the Newspaper Industry in Transition” is available at http://www.crl.edu/sites/default/files/attachments/pages/LCreport_final.pdf.

strategies for collecting and preserving digital news, both web and broadcast, on a national basis. It was clear to all who attended that devising effective strategies for preserving news in the electronic environment will require an understanding of the “lifecycle” of news content.

Prompted by the workshop discussions, CRL set about examining, analyzing, and documenting the changing flow of news information, content, and data in one sector of the increasingly diverse news ecosystem: newspaper journalism. CRL focused on the lifecycle of news, from news-gathering and sourcing of content, through editing and processing, to distribution to end users.

The ultimate goal of our study was to pinpoint the potential “high-impact point of entry” in this lifecycle where libraries and other memory organizations might intervene to capture critical news content and related metadata in a way that would ensure the long-term survival and accessibility of the content for historical purposes.

1. The CRL Study

Our study particularly addressed the following points of interest:

1. *The nature of the electronic facsimile or replica paper:* LC was particularly interested in determining how one particular news format, the “electronic facsimile” of the printed newspaper, i.e., the PDF and other page-image files for a print edition delivered to printers and aggregators like ProQuest and PressDisplay, might be exploited for legal deposit and archiving purposes.

2. *The relationship between Web and print news "coverage”:* LC wanted to understand how content posted to a given newspaper’s Web site compared with the contents of the print edition.

3. *Technical formatting and delivery of electronic news output:* To what extent are there emerging industry technical standards for formatting, managing and/or disseminating news content? And what is the range of current practices for formatting and encoding news for distribution in print and on the Web?

Four U.S. newspapers provided case studies for the project: *The Arizona Republic, Wisconsin State Journal*, *The Chicago Tribune and the Seattle Post-Intelligencer* (since 2008, *seattlepi.com*). The case study titles were chosen to provide a range of types of newspapers that represented a broad segment of the U.S. newspaper industry and its publishers. The report also drew from previous CRL analyses of three other news organizations: the New York Times Company, Investor's Business Daily, and the Associated Press.

For purposes of the report we identified the **basic stages of this lifecycle as:**

A. *Sourcing:* the gathering of news information and content by the news organization from those who create, report, and/or own that information and content;

B. *Editing or production:* the editing, processing, and enhancement of news content and information, and preparation of news products.

C. *Distribution:* dissemination and exposure of news content, and products and derivatives thereof, through print and online media.

Essential to each stage of this lifecycle **are automated systems deployed to produce, modify, and annotate content and prepare the products of the news publishing process.**

The major systems involved are editorial, digital asset management or archives, pagination, and Web production systems. In addition we looked at third-party providers who, working in tandem with news publishers, employ their own systems to produce and deliver data and other news content through the publisher's Web and print channels.

2. Our Findings

I will limit my remarks here to the **three factors that Web news distribution introduces that have profound implications for the preservation of digital news.**

First, at almost every point in the digital workflow **news content is annotated and enhanced by creators, publishers and the technologies they employ with metadata** that provides valuable information on source, authorship, rights, version, subjects, and technical properties. That metadata is critical to the creation of news products, such as websites, print editions, RSS feeds, and other outputs, and to the management and exchange of news content among systems and news organizations.

Industry-standard exchange formats such as IPTC 7901, ANPA and NITF, originally developed for structuring newspaper text content for transmission and used in marking up simple news items and article texts, **provide the ability to embed a tremendous amount of information in a piece of electronic content.**

These exchange formats are descendant from the codes developed in the last century by the Associated Press for delivery of text via dedicated newswire networks. NewsML™ and APPL (Associated Press Publishing Language) are newer, media-type agnostic news exchange standards that enable news organizations to provide a wealth of descriptive, structural and administrative information on transmitted news objects (e.g., reports, articles, photographs) that can be read by automated production systems at newspapers and contractors within the publisher's network. They enable production systems, for example, to manage articles, photographs and other content over time by providing information on rights status (publishable, embargoed, etc.) and information on authorship and copyrights (source, credit line, terms of use, etc.). They also enable news agencies and publishers to generate the same text in multiple languages; video clips in different formats; and the same photograph in different resolutions.

Similarly, a news photograph transmitted from a photo agency or a photographer to a given newspaper will, at minimum, normally include information about its authorship, source, technical characteristics, subject matter, urgency, and general terms of use. It might also include, however, an abstract or descriptive information, version history, and related media objects.

In short, an enormous amount of information about a given news article, photograph or other feature is attached to the item and maintained within the editorial systems of the publishers, wire services and agencies.

Most of this information is stripped from the item when it is “published” or output to printers and aggregators in PDF and ASCII formats, and to the Web in HTML.

As a result, preserving the printed edition, the electronic facsimile alone necessarily misses information that could be quite useful to researchers.

Moreover, the contents of the print editions are gradually but inexorably becoming a **smaller portion of the total daily news output of the publishers.** There is a great deal of Web-only news reporting and writing coming from the newspaper publishers, and because of the frequency of updates to newspaper websites, the number of electronically “published” versions of a given article or feature has multiplied exponentially.

Finally, **production of the e-facsimiles may well decline** as consumers become more accustomed to reading on the Web and on personal devices. Tablet devices, in particular, are rapidly gaining favor as a means of news consumption. Together with the decline of print circulation, this could lead to the discontinuation of PDF production by publishers, and render the effectiveness of a PDF acquisition program short-lived.

Second, much of the content that appears on the web site of a news organization is sourced from other providers, and is not born in the editorial processes and systems of the news organization. Much content, including financial data, advertising, reader comments, and so forth, is rather piped directly to the Web from specialized third-party providers. Those providers alone control the flow of this content, and they alone hold the rights to its use. Their content never actually “resides” within the editorial space or control of the news publisher. This means that most news organizations are no longer in a position to provide libraries with the entire content of their news products. Therefore, legal deposit of a complete electronic “edition,” of their publication, including third-party electronic content, would not be within the power of most publishers to provide.

The **third factor** has implications for libraries that endeavor to harvest the websites of news organizations as a form of preservation. Although the dynamic quality of digital media is well known, we learned that the **contents of the web edition of a newspaper change with a frequency that is not only erratic and unpredictable, but which varies from one type of content to another**.

A monitoring study of the *Chicago Tribune* website conducted over 30 days by Kalev Leetaru of the University of Illinois, found that the rate of change even in the text content alone of the Tribune's site was so frequent and irregular that to ensure that every new article posted to the site was harvested, an archivist would "have to monitor 105 main topic pages on the site every few hours or risk losing new articles on a news-heavy day." The study monitored 41,220 Tribune URLs linked from Tribune gateway pages during a 34-day period. The histogram indicates that only a small percentage of those URLs, fewer than 3,000 (7%), persisted for more than 36 hours.

The study showed that news articles are updated more frequently than feature articles, for instance. And financial data is updated more rapidly than news articles, but only during hours of market activity. This means that, unlike with printed newspapers, there is no single authoritative web edition, or "edition of record", of a given day's news site.

This fact makes newspaper websites singularly difficult, if not impossible, to archive by the current web harvesting crawlers.

It is not surprising then that many of the existing Web harvesters only incompletely capture the full contents and functionality of newspaper web sites. Harvesters are unable to keep up with the rapid pace of content updates on the major news sites like the New York Times and Chicago Tribune, a pace that only accelerates with each passing year. The Internet Archive's harvest of the New York Times between December 12, 1998 and September 05, 2007, as represented in the Wayback Machine, holds only about 352 "issues" of http://www.nytimes.com, the landing page of *The New York Times*. In these, much of the non-text content (images, graphics) is missing and many links (like AP and Reuters feed) are not functional. The Library of Congress's recent crawls of the Detroit Free Press improves upon the Internet Archive's results, but still is hampered by similar limitations.

Moreover, the pay walls now being erected by the News Corporation, New York Times, Wall Street Journal, and others present a new obstacle to harvesting news from the Web sites of publishers. And the subscription services planned by Apple (per the App Store) and Google (One Pass) may put yet another layer of defense between the harvester and targeted news content.

Therefore, we believe that the current preservation methods, archiving PDF files and newspaper websites, have inherent limitations. We believe that neither of these approaches to preserving news will serve future researchers well.

3 Some Suggested Strategies

So what are libraries to do? The CRL report suggests several strategies, and CRL and North American research libraries have begun to pursue on some of them.

3.1 National site licenses

In the past, national libraries like the British Library and Library of Congress, and major library organizations like CRL, together formed a key link in the newspaper supply chain for the major academic and independent research libraries. This role was supported by the national legal deposit systems, CRL and LC microfilming operations, and by collecting through overseas field offices. Today, those activities, unfortunately, are no longer effective, defeated by economic, technical and policy changes: funding shortfalls, the rapidly rising costs--and risks -- of maintaining foreign pipelines, as well as changes in news distribution technologies.

With the convergence of the "vertical" media, libraries might consider organizing their news collecting and preservation efforts around some of the major media organizations, such as the New York Times, Associated Press, and News Corporation; or sectors (financial information, advertising, news cooperatives), rather than around formats (newspapers, serials, broadcast).

In the U.S. and Great Britain, content production and management activities are increasingly centralized, as a result of the growing cost of content management and consolidation in the news industry in the wake of deregulation of media ownership. The same factors are driving the merging of editorial and content management systems and reliance upon data centers to process and store content shared across a given media organization's several platforms. This consolidation promotes uniformity of practice, and might thus present an opportunity for libraries. Could libraries, acting together, influence the major media organizations to manage their content in ways that would better serve current academic and policy researchers and future historians?

This approach would involve working not with local news organizations but with the large parent companies that increasingly control the content of their local newspaper properties. These include media titans like Gannett, the News Corporation, Tribune Company, and McClatchy. As some of these also control and create content for radio and television outlets, dealing with them might yield benefits for collections in the broadcast arena as well.

U.S. and Canadian libraries working together through the agency of CRL, are now exploring how we might secure preservation services and rights through negotiated licenses of digital news databases. The terms of these licenses will be structured to guarantee uninterrupted, long-term access to the news content for research purposes, provided by the publishers and/or the major aggregators. Such licenses will build upon current library subscriptions to paid news services like the Wall Street Journal and to text aggregator services such as LexisNexis, Factiva, and Access World News. Such licenses will have to be iron-clad and ensure specific rights and privileges with regard to use of the news content over the long term. Concrete assurances for such persistent access to the content may be able to be provided either through a trusted third-party dark archive (along the lines of Portico or CLOCKSS), or through rigorous data maintenance measures taken by the publishers and aggregators themselves and periodically audited. (CRL is currently assessing the self-archiving systems of ProQuest and Readex.)

Such an arrangement would have to be negotiated on behalf of the U.S. research libraries with aggregators like NewsBank, Dow Jones, and LexisNexis, or with the media organizations themselves. The latter are yearly becoming fewer and fewer in number. And the largest organizations, like the News Corporation, Bloomberg, and Associated Press, are producing news for multiple delivery platforms: print, Web and broadcast. Therefore dealing with those organizations could conceivably satisfy research libraries' needs for news from all three types of platforms.

This approach will indeed entail a substantial investment by CRL and its community. But that investment would pale in comparison with the cost of replicating the functionality of the many systems required to store and manage electronic news as it is produced today. A side benefit of such an arrangement might be the exchange of technology, preservation, and subject matter expertise and knowledge between the media/publishing world and the CRL community.

The scale of such an investment, moreover, might well purchase for research libraries some influence on the production and content management practices of the news organizations. Representing a significant sector of the media organizations' customer base could position CRL to force standardization and uniformity that would reduce the future costs of its "taking custody" of the content and necessary enabling systems in the future. Under this arrangement the research library community would not "own" the content, but would exercise a measure of control over it.

3.2 The value of documenting news production and distribution systems and processes

We sometimes forget the role libraries have played in the past as authenticators of evidence for larger societal purposes. Documenting and mapping the processes involved in producing and exchanging news content has a forensic value. Enlarging our understanding of these processes could help preserve for future researchers and other users the ability to analyze evidence and excavate information produced by discarded systems. I use the term "other users" here significantly. This may be particularly important in the arena of law and government, which have distinct needs regarding the preservation and presentation of evidence, and where newspapers have always played an important role in documenting contested events and actions. CRL recently undertook a study for the John and Catherine T. MacArthur Foundation, on the production and use of electronic evidence of human rights violations. In many recent world events, the Arab Spring included, digital video, blogs, FaceBook postings, and Twitter feed featured prominently. Videos posted to YouTube, for example, provided a visual record of the events in Tahrir Square and post-election violence in Iran and Kenya. We wondered how well that kind of evidence would hold up for purposes of societal record in the long term; and in the near-term for purposes of prosecutions and other legal proceedings.

As part of that study we commissioned two papers: the first on emerging practices with regard to admission of electronic documentation in cases in domestic and international courts; and the second on the principles governing such practices. The authors of both reports stressed the importance of technological expertise in demonstrating the authenticity of documentary evidence, and its chain of custody.

In the past, precedent and the laws of evidence established what forms of documentation could be brought to bear in civil and criminal court actions. Understanding and documenting the processes by which news is gathered, edited, transformed, and distributed today will be essential to establishing the chain of custody and authenticity of news content in the future, for scholars and jurists. Knowing how the content of a given article or photograph, for example, changed in passing from one production or storage system to the next, or was annotated and enhanced with metadata in the process of editorial work, could provide insights on its value or even inform development of a news preservation strategy. We fear that not enough is known about the current methods of producing and distributing digital news to ensure the integrity or traceability of that chain of custody. CRL's mapping was a first attempt at that kind of analysis.

We also need to look closely at the relationship between the content created by newspapers and the proprietary data and other content provided by major third party organizations, such as weather data gathered and preserved by the National Center for Atmospheric Research and National Weather Service; financial data produced and distributed by Bloomberg; and public opinion data from Nielsen, Pew, and others. Understanding how the databases and the digital asset management systems of the news giants like Associated Press and Gannett function and interact will be important in evaluating and interpreting the electronic news that survives. **Aside from libraries, no other institutions are likely to play this role in today's society.**

3.3 Understanding new and emerging research needs

Before we have a strategy, however, it will be necessary to examine the underlying goals of our news preservation program. Most library preservation and acquisition policies are based on certain premises that were valid during the print era, before the radical changes in information production and consumption brought about by digital technologies. Those premises need to be reexamined in the light of today's technologies. We may be basing our policies – and investments – on false assumptions about the practices and needs of contemporary and future scholars and researchers.

For example, a commonly stated library goal is to preserve today's news in a form in which future researchers not only can recover the content of the news but can understand how contemporary citizens perceived and experienced the news. In the age of dynamic media, this is a tall order The proliferation of devices for accessing the same news content (mobile phones, tablets, PCs, e-readers, etc.) and variety of applications used for presenting that news (RSS feeds, news readers, iPhone apps) atomizes the user's "experience" of electronic news into a million variations. In addition, the online transaction between the producers and consumers of today's news involves customization of the content to individual user traits. As “real-time analytics" are built into the technologies for news distribution increasingly shape the content of advertising and news, it is probably unrealistic to expect to be able to reconstruct all types of news consumption experiences -- or even a typical experience -- in the future.

The needs of today's users have also changed radically in recent years. Financial, legal, public policy and academic researchers increasingly employ computers to locate – and in some cases even to interpret -- information in large bodies of news content. Witness the financial industry's large recent investment in news mining engines.

The value of metadata for economic and policy researchers might rival that of the content itself, e.g., for generating new information and findings about people, organizations, subjects and places. Rich metadata added by publishers and aggregators enables computer-assisted quantitative and qualitative analysis across large bodies of text and media content that is not possible using the published content alone.

Therefore it is safe to assume that researchers will be mining Web news for different types of information than they sought from newsprint. To construct effective preservation strategies we simply do not know enough about those uses.

4. Moving Forward

I have offered here suggestions for how libraries might help to ensure that the digital journalistic record is preserved for future scholars. Based on the findings of the CRL study I have described in broad terms what an effective effort to preserve digital news might "look like." These ideas need to be explored in further depth. But CRL has chosen to move ahead in the meanwhile down this path, to support advanced research at U.S. and Canadian libraries. The ICON project, a longstanding news preservation projected under the auspices of CRL Global Resources, will be the locus of this planning and strategic action.

There is some urgency to our moving ahead on an international basis. We are now midway through the second decade in which the Web has been a major venue for news "publication," without a viable comprehensive plan for the systematic capture of digital news. We encourage national libraries and the research libraries of the world to join with us, to act boldly and decisively, to begin to close the widening gap in the historical record.   For the present situation is a recipe for failure – and irrelevance.